EDI: ELECTRONIC DATA INTERCHANGE
FOR STATISTICAL DATACOLLECTION AND DISSEMINATION

W.J. Keller, Statistics Netherlands, The Netherlands

ABSTRACT

In this paper, we will present some experiences in the Netherlands with EDI for statistical datacollection and dissemination. We will consider the changes to be made for large scale EDI datacollection. We will argue that EDI demands a dramatic redesign of the way we collect and process statistical information, but the rewards in terms of response burden, quality and efficiency might be well worth it.

We will also discuss some projects at Statistics Netherlands dealing with EDI for statistical dissemination. We will cover Statline (our statistical database with a traditional on-line querytool on a remote DOS client) and its new experimental version, with a so-called dynamic WEB on Internet. We will argue that the Internet not only provides great opportunities but also great challenges for statisticians.

We will focus, besides on the technological aspects, on the conceptual and organisational implications EDI for Datacollection and dissemination.

Keywords: Official Statistics, Datacollection, Dissemination, EDI, Internet, Meta-Information

1. OFFICIAL STATISTICS

National Statistical Institutes (NSI’s) are confronted with several strategic issues resulting from new demands from our customers as well as new developments in Information Technology (IT). Efficiency and market-orientation are the key-words now. We need to produce at lower costs. Furthermore we need to lower the costs we inflict upon our suppliers of data. The outcome should be a product that, although not actually sold on a market, our clients eventually want. Furthermore we are confronted with new developments in IT. They will give us the opportunities to construct the necessary tools to meet the new demands. In a situation like this a NSI needs to make the right strategic choices.

The statistical production process is influenced by the growing demands of our clients (output) and of our respondents (input). Concerning our output we see a demand for a better access, preferably electronically, and a greater user-friendliness. One particular aspect is a demand for an improvement of the coherence of the totality of the information we offer. W.r.t. our input, there is a strong political demand for a decrease in the respondent burden as a part of alleviating the administrative burden of enterprises. NSI’s like Statistics Netherlands sends out a million questionnaires to enterprises and other institutions per annum. Large and medium-sized enterprises may receive as many as 50 questionnaires per year, including repetitive monthly and quarterly surveys. The conclusion is clear: NSI’s have “to fight the form-filling burden”. Furthermore, budgets are shrinking so there is a demand for higher efficiency and higher productivity.

2. INFORMATION TECHNOLOGY (IT)

We are also blessed with new IT developments: the technology push. These developments give us new technical possibilities, the means to construct new tools for our production process. We see large improvements in the possibilities of data processing, data storage and data transmission. The last aspect will probably have the most striking influence on our work: the electronic data interchange (EDI) between our respondents and the NSI on the one hand and the EDI between the NSI and its clients on the other.

But these new IT developments also create their own demand outside the NSI. The new technology will be used anywhere. Our suppliers of data will use it. Our clients will use it. They will no longer be satisfied to communicate with us in the old way, that is on paper. Our suppliers produce their data by electronic means and will want to use those means to deliver those data directly to us in order to minimise their own costs. Our clients process our data by electronic means. They will demand to be able to select and receive those data with the tools that IT has to offer.

These factors lead to the conclusion that the NSI will have to make those strategic choices in its production process that make the best use of the possibilities IT has to offer. The potential of IT will affect all aspects of our production process. To describe them let us discern, within this production process, three stages. The input-phase is where the data are collected in contact with the respondents. In the throughput-phase these data are processed to produce the information with the characteristics we are actually looking for. In the output-phase this information is offered to and disseminated among our clients. In this paper, we will concentrate on the external effects, so on the input and output phase.

Since the output phase defines our product and relates to our customers, we will consider it first. Here the new developments probably get the most attention from the public. We see the new media by which information can be presented to its users. Paper publications may continue to play their role but especially the more professional user will want to select and receive his data by electronic means. NSI's are producing or developing those means: data on CD-ROM, data on Internet. More important and maybe more difficult is the way data should be presented with those new media. The amount of information will be much larger than we had in our paper publications. At that point the management of the meta-information becomes crucial.

For this purpose NSI's are developing statistical databases, intended for end-users, giving access to "all" our data. As could be expected, structuring those data is the main problem. At the same time we are confronted with lacking coherence due to lacking statistical co-ordination. Output databases are intended to play a key-role in the dissemination process of our data. In the future, the strategic choice should been made that we aim for such a structure wherein all publications and all other dissemination of data goes through the database. Section 3 and following will discuss the output strategies in more detail.

Besides the output side, especially the data-collection will be re-ordered. Besides the use of Computer Aided Interviewing (CAI) for household surveys, also EDI will play a key role here, in particular for establishment surveys. Because of the nature of EDI, no longer the demand for information (i.e. each statistical survey), but the supply of information from our respondents will dictate the organisation there. And since most information from our respondents indirectly comes from electronic data-sources like bookkeeping systems and registers, we should concentrate on these sources. If possible, each source should be tapped electronically once and completely, using EDI, for any possible use within the NSI. The collection is technically and conceptually adapted to that source. In section 7 below we will elaborate on this view.

In this paper we focus on EDI at the output and input side of the statistical process. We start with the output side.

3. DISSEMINATION

At present, most statistical agencies provide aggregated statistical information in various ways, but dominantly in printed form (on paper). Because printing is a relative cumbersome and expensive way of dissemination, more and more people are looking at the electronic highway (a.k.a. the Internet) as a cheap and easy way to disseminate statistical information. This paper focuses on the impact this trend has on official statistics. We will argue that besides the technological dimension of publishing on the Internet, the main problems will be conceptual, i.e. those of statistical co-ordination and integration.

In this paper we will discuss some projects at Statistics Netherlands (SN) dealing with EDI for dissemination. We will cover Statline (our statistical database with a traditional on-line query tool on a remote DOS client) and its new experimental version, Statline-WITCH, with so-called dynamic Web pages on Internet. We will argue that by combining the ease of use and ease of access of the Internet with the multi-dimensional database systems found in statistics, great opportunities for statistical dissemination will arrive.

Presently, our publications take many different shapes: printed paper, floppy disks, faxes and CD-ROM’s, automatic and human voice response, press release, videotex, etc. Behind all these different media there is (aggregated) statistical information, often in machine readable form e.g. as the output of survey processing systems. Needed is a "one-stop" dissemination database situated between the internal processing systems and the outside world, capable of producing many different media from one source, in a consistent, timely and efficient way. Besides on-line access to the database, such a database system could also automatically provide the information for other media, such as floppies, email subscriptions, faxes and CD ROM publications. But one of the most important objective of such a system is to provide easier on-line access by our customers to the wealth of information at statistical bureau’s. It is our opinion that in this respect the Internet will play a very important role in the near future.

With the rise of the graphical browsers (Mosaic, Netscape) on the Internet, the net has grown immensely during the last year. Within months, nearly every respectable company has set up its own so-called “Web-server” on the World Wide Web (WWW). The net, with its sky-rocketing popularity and therefore great infrastructure, is already connecting tens of millions people all over the world, with access becoming easier and bandwidth nearly free (in Holland, the Internet, at 28 kbps speed, will be a local phone call away for nearly everyone at the end of 1995). It allows statisticians not only to collect information more efficiently (see our paper on EDI), but also to disseminate aggregate statistics more efficiently, with a marginal reproduction and distribution price close to zero. (There are already 27 000 free Internet subscribers to David Letterman Top Ten Listserver: imagine such a circulation to our press releases !)

At present, several statistical institutes publish information on the net through the WWW. Well-known Web-servers are those from the US Census Bureau, Statistics Canada, Eurostat and SN, to name a few. Everyone with an Internet connection and a browser like Netscape can visit these servers from all over the world. Most of the material published on these Web's, however, is not really statistical information, but lists of publications, press releases, and general information for the public. The limited amount of truly statistical figures is often presented in a documentary way, i.e. as electronic copies of the printed pages from traditional publications.

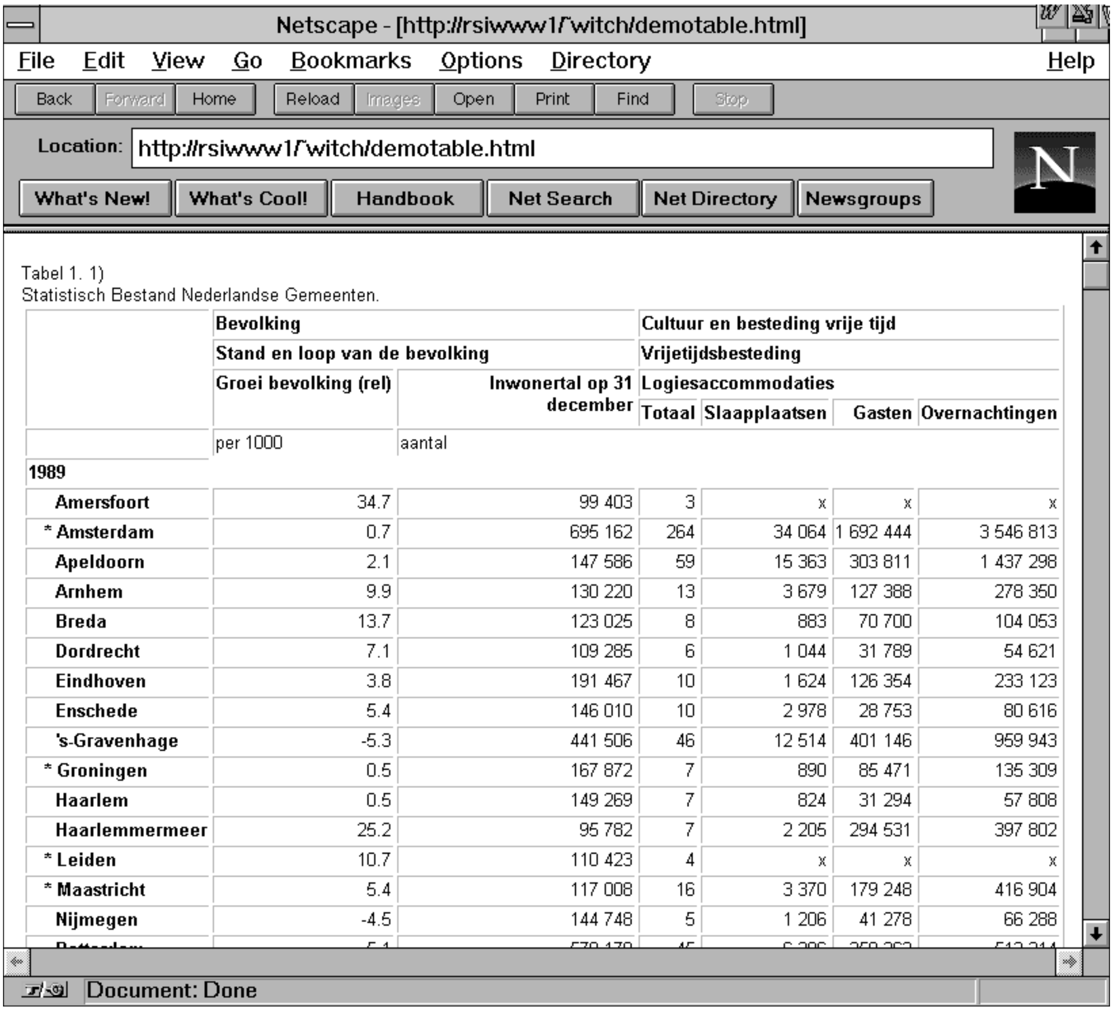
This approach, which is typical for so-called *static Web pages*, makes it difficult to manipulate statistical figures as structured information, since the user only has access to documents, i.e. (formatted) text. What is really needed is access (through the Internet) to a real *database*, encompassing various statistical sources in an integrated system. Once our statistical information is available in a structured, machine readable way, we can manipulate it and present it in any form, including unstructured (like a text document) and structured (e.g. like a spreadsheet). This structured database approach is also necessary in order to be able to provide better co-ordinated and integrated statistical information.

An example of a statistical database is Statline, from SN. Statline is based on the client/server concept, where the front end (running on a PC, possibly outside SN) is separated from the back-end (the database server, located at SN). Front end and back-end are presently connected through traditional datacommunication facilities like Local Area Networks (LAN's) internally or simple asynchronous lines (using telephone lines and modems) externally. In order to optimize the performance of its multi-dimensional database (see section 4), Statline uses a proprietary, non-relational database design based upon indexed files. The DOS-based front-end uses a user-friendly window/mouse desktop metaphor where the results of searches are displayed in a type of multi-dimensional spreadsheet, with additional graphical views, including thematic maps. The Statline front-end is the same as the software we use as interface to our floppy disk (or CD-ROM) -based publications. Presently, Statline does not use the Internet, but this will change when we introduce the concept of *dynamic Web pages*.

4. DYNAMIC WEB PAGES: COMBINING INTERNET WITH DATABASES

As discussed above, the statistical information found on ordinary Web pages on the Internet is difficult to manipulate in a structured way, in view of the documentary (non-numerical) character of a Web page. Also, each Web page is static in nature, i.e. we have to prepare each page beforehand by storing its (documentary) image on the Web server. Wouldn't it be great to combine the power of on-line databases, like our Statline database, with the ease of use and access of the World Wide Web? This is where the so-called *dynamic Web page* enters the picture. The idea is to use browsers like Netscape as front end to systems like Statline. Each time a user requests data, a special interface, called WITCH, translates the request to the Statline format and generates a Web page on-the-fly to present the result from Statline to the user.

An example of a WITCH generated Web page, using Netscape 1.1 with HTML3 table-support, is shown below.


Netscape - [http://rsiwww1/'witch/demotable.html]
File Edit View Go Bookmarks Options Directory Help
Back Forward Home Reload Images Open Print Find Stop
Location: http://rsiwww1/'witch/demotable.html
What's New! What's Cool! Handbook Net Search Net Directory Newsgroups
Document: Done

Tabel 1. 1)
Statistisch Bestand Nederlandse Gemeenten.

	Bevolking		Cultuur en besteding vrije tijd			
	Stand en loop van de bevolking		Vrijetijdsbesteding			
	Groei bevolking (rel)	Inwonertal op 31 december	Logiesaccommodaties			
			Totaal	Slaapplaatsen	Gasten	Overnachtingen
	per 1000	aantal				
1989						
Amersfoort	34.7	99 403	3	x	x	x
* Amsterdam	0.7	695 162	264	34 064	1 692 444	3 546 813
Apeldoorn	2.1	147 586	59	15 363	303 811	1 437 298
Arnhem	9.9	130 220	13	3 679	127 388	278 350
Breda	13.7	123 025	8	883	70 700	104 053
Dordrecht	7.1	109 285	6	1 044	31 789	54 621
Eindhoven	3.8	191 467	10	1 624	126 354	233 123
Enschede	5.4	146 010	10	2 978	28 753	80 616
's-Gravenhage	-5.3	441 506	46	12 514	401 146	959 943
* Groningen	0.5	167 872	7	890	85 471	135 309
Haarlem	0.5	149 269	7	824	31 294	57 808
Haarlemmermeer	25.2	95 782	7	2 205	294 531	397 802
* Leiden	10.7	110 423	4	x	x	x
* Maastricht	5.4	117 008	16	3 370	179 248	416 904
Nijmegen	-4.5	144 748	5	1 206	41 278	66 288

By using a Web-browser as front-end to a database with structured information, other Web-tools also become available. For example, besides presentation in a Web format, we can also download information or use other “viewers” in the browser, e.g. to see spreadsheets, graphs, or maps from the net. WITCH will not only generate dynamic Web pages but other formats like spreadsheets as well. In this way, the user can save information in a structured format in order to manipulate the data later.

The advantages of this approach are several: first, we don’t have to build our own front end tool, like we did with Statline for DOS. Anyone with a decent Web-browser can access Statline, wherever in the world. Second, by using the commonly available Web browser, Statline becomes immediately available on different platforms (Windows, Mac, UNIX). Third, the user does not has to learn a new interface, once the Web browser is known. Finally, we can use the Internet as communication medium, with all its advantages: high bandwidth (28.8 Kbps by modem or even better in case of ISDN or T1 links) and great accessibility (as said before, in the Netherlands the Internet will be a local phone call away for nearly everyone at the end of 1995).

With millions of potential users being able to access our on-line statistical databases over the Internet, new challenges arrive. The biggest concern, as we see it, is the statistical co-ordination of the information we provide.

5. STATISTICAL CO-ORDINATION

Most statistical bureau’s provides hundreds of different statistical publications from several hundreds of surveys. All this amounts to million of figures, thousands of tabulations, and many, many different sources of information. But except for some special publications (like the National Accounts), each publication only deals with a very specific topic, and users are confronted with an inaccessible "gold mine of information" with many, many different faces. Someone being interested in, say, automobiles, has to look in more than a dozen publications to get a total picture, encompassing the production of cars, the exports and imports, the use (in time and mileage), the energy consumption, traffic accidents, the environmental effects, etc. Finding all this information can be laborious en troublesome, especially since each statistical department focuses only on their topics and publications. At the same time, NSI’s sell only a very limited number of copies of each individual publication, often without recovering the full dissemination costs, let alone the collection costs. And finally, while users appreciate our impartiality and accuracy, they complain about the lack of timeliness of our statistical information.

If all available statistical information is placed on the Internet, free of charge, millions of users can and will access it. Compare this with the hundreds of users reading our printed publications. However, not only the implications in terms of distribution are mind dazzling, also the conceptual implications will be great and probably very problematic. Why? Since with such a unlimited access to all statistical information, users will ask much better access paths (with search by keyword and multi-dimensional queries, on time, branch, region, etc.. on top). And then, after we have provided these tools, they will find out that our information is not always co-ordinated, let alone integrated. Inconsistencies, buried in hundreds of different paper publications, will become visible on the net, and users will start asking questions: not only for more, but also for better co-ordinated and better structured information.

One answer to this demand for better co-ordinated data is the systems approach, like National Accounts. Another, less ambitious goal, is to co-ordinate the classifications, domains and definitions used in different statistical publications. This is the philosophy behind a new database approach, based on the concept of multi-dimensional tables or *cubicles.*

As in Computer Assisted Interviewing (CAI) systems, we can distinguish between the data itself and its description, the so-called metadata. While the CAI systems focus on the individual data and metadata processed in the data collection and editing stages, in the dissemination database we focus on the aggregated data and its metadata. This metadata comprises both the syntax (format) and semantics of the data (the definitions of the published variables, like the definition of “number of employees”), as a description of the survey itself, the sources and how the items are derived. The first step to co-ordinate statistical publications is to standardize the definitions of the variables used.

Each item (e.g. number of employees) is often available for different domains, defined by crossings of discrete, categorical variables, like sector, region or time. An other important mechanism to co-ordinate the dissemination of statistical data is the standardization of these categorical variables, leading to classifications e.g. for branches of industries, commodities, regions, etc. The basic representation of information used in such a database is therefore the multi-dimensional matrix (sometimes called "cubicle") where one dimension reflects the different variables (e.g. number of employees, profit, prices), a second one the (discrete) time axis (e.g. years and months), while other dimensions correspond to various classifications (industries, commodities, regions, etc.). The items inside the matrix reflect the measurements ("number of") on a certain variable ("employees") in the domain defined by the crossing of the categories on the other axis ("in industry x in region y at time t"). Often, categories are classified into different systems of detail (e.g. a n-digit industries classification, with n=1..9) which are often (but not always!) hierarchical to one another, resulting into levels of classification.

Metadata (descriptions) in this database of cubicles can refer to the total matrix, to the axes and their variables and categories, and to the individual items inside the matrix. Particular problems of metadata arise when the definitions of certain categories (like regions, industries, commodities) change over domains, in particular over time. As example, take a region like a municipality. Not only the number of inhabitants in Amsterdam in 1980 is different from 1990, but also the definition of Amsterdam itself differs between the two years (e.g. because of border corrections)! Similar problems arise when certain items are only available for certain categories or classification levels, making comparison of different items in various domains sometimes impossible.

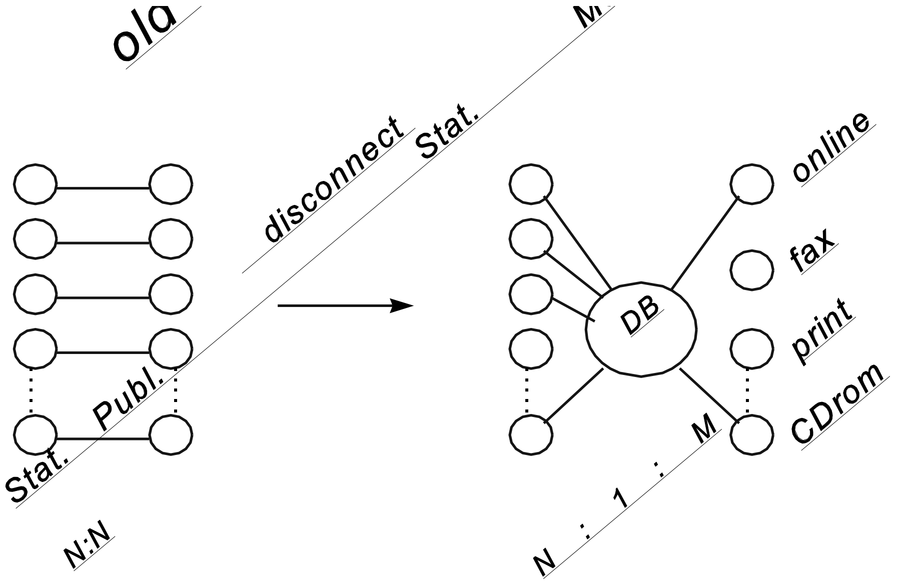
As explained above, in statistical databases the most important type is the multi-dimensional object, or *cubicle*. A statistical database will contain many, many different cubicles, which might all share similar classifications along some of their axis. Besides these multi-dimensional objects, also simple ("flat") two-dimensional cross-tabulations, as shown in most traditional statistical publications, have to be stored and presented in the database, as well as (one dimensional) text objects like press releases. All this information is documented (metadata) inside the database on various levels (from the total object down to the individual items or cells). A classification of database objects into the well-known statistical domains (like economic, social and demographic statistics), and classifications thereof (production, environment, labor-market, well-being, etc.), make navigating through this immense database of information more feasible. A very strong tool in finding the information needed is a database-wide keyword (thesaurus) system, which allows the user to quickly allocate the right object.

In several countries, statistical databases based on this concept of cubicles are presently being used or under construction. Well-known systems include PC-Axis from Statistics Sweden, the ABS-Database from the Australian Bureau of Statistics and the above mentioned Statline database from SN. Statline is both an internal SN system as well as an open on-line database system, available for customers outside SN. Inside SN, it will be used to "drive" the publication of different media (paper, floppy, videotext, etc.) as much as possible without any human intervention (e.g. with fully automatic composition of printed material). Statline is also used for all our internal inquiries, e.g. for customer support, and to provide customers with snapshots from Statline on a incidental or regular basis (e.g. with automatic fax/email subscription on "hot" figures). It will also, and possibly most important, be used as a vehicle to standardize (and therefore co-ordinate) all our aggregated statistical information, including our metadata.
Outside SN, Statline provides a direct connection for our large accounts to the wealth of information Statistics Netherlands provide. By combining the WITCH interface, using the idea of dynamic Web pages, we will allow for easy access to Statline over the Internet to many other customers.

6. DISSEMINATION: CONCLUSIONS

With the spectacular rise in the use of the Internet world-wide, electronic publishing quickly becomes a reality. The Internet and in particular the WWW (World Wide Web) not only provides great ease of use and ease of access to an immense universe of information, it also provides great challenges to statisticians. Should we simply put all our paper-based publications in electronic form on a Web server, using the same document form as we did in printing? Or should we make our statistical information available in a more structured way? We think that the technology of the so-called dynamic Web page as a front-end to a database with statistical figures will be a better solution than static Web pages, which in some way just replicates the paper metaphor.

More in general, once statistical information is available in a structured, machine-readable format like Statline, we can present it in any form by just using interfaces like WITCH. From such a database, not only Web pages can be generated on the fly, but also fax/email messages, press releases, databases on CD-ROM and even old-fashioned printed output in a completely automated way. Of course, not only the data itself but also the metadata should be machine readable, including the syntax (format) and the semantics (content) of the data. Once this is achieved, we can easily exchange information between statisticians using standard export formats like Eurostat's GESMES, like we now use WordPerfect exports from a MS-Word document. All it takes is a structured, machine readable and documented form of storage of all statistical information.



Even if all our information is available on-line, machine readable, and well documented with a lot of metadata, users will start complaining again as soon as inconsistencies between electronic publications become visible. Then, we will need some way of statistical co-ordination and integration, like we did with the National Accounts, but now on a larger scheme. Trying to integrate as many publications as possible in a limited number of cubicles, might be a first step into the right direction. To do so will ask for a great effort in streamlining statistical definitions and classifications. In the end, the conceptual problems might overshadow the technical ones.

Finally, there is the interesting aspect of cost and price on the Internet. Assuming that statistical information itself is a public good, statisticians are often pricing statistical information only according to the marginal costs of reproduction and distribution. Reproduction and distribution on the Internet is essentially free. So, one might wonder what to do once we are able to put *all* our gigabytes of available statistical information on the Internet. Should it be free of charge or not? This topic has resulted in lively debates at SN. So besides the technical and conceptual problems, the Internet also raises great strategically issues.

7. RESTRUCTURING THE PRODUCTION PROCESS: DATACOLLECTION

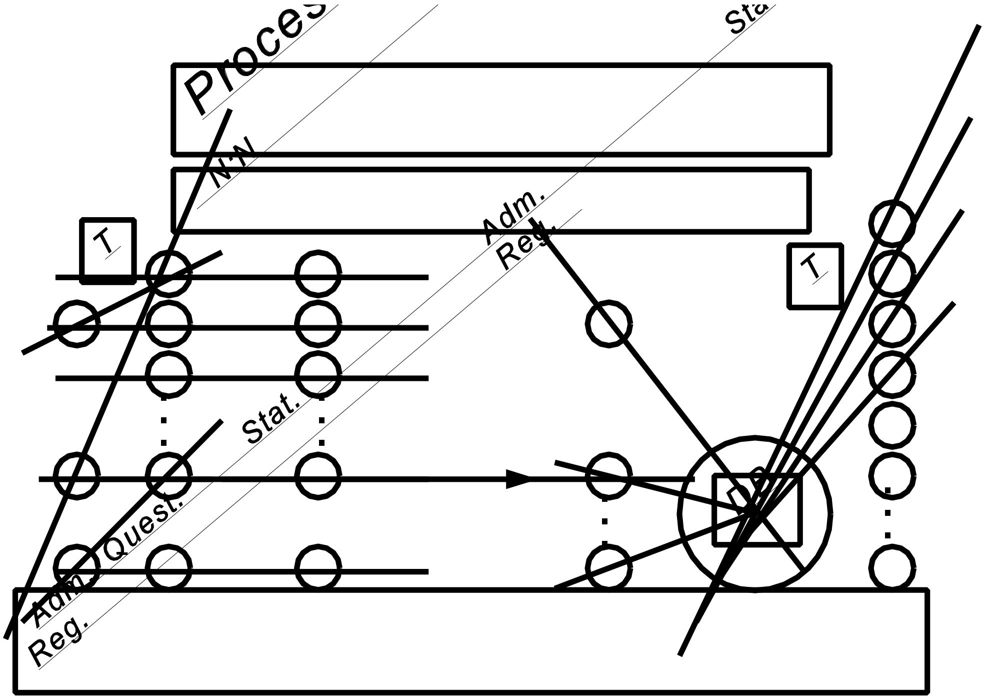
In the previous section we described the strategic choices to be made regarding the output phases of our production process. Those choices go further than just the development of a new tool. They will affect the structure of the production process itself. One should be prepared to take those consequences as well. The present or the "old" way the production process is structured is along the lines of the individual statistics. For each statistic - an end-product - a new questionnaire is designed, respondents are selected, data are processed and a publication is made. Especially on the input side this is very inefficient.

Let us first consider data-collection among individuals and households. It is not saying too much when we state that a major step forward has already been taken at Statistics Netherlands. We have introduced all kinds of Computer Aided Interviewing (CAI) and developed BLAISE to do so. (Needless to say that BLAISE does more than develop and present electronic questionnaires.) The gains of these developments was mainly in terms of an increase in productivity or efficiency. The number of staff needed for coding, data entry and checking decreased dramatically. This efficiency also shows itself in the much faster production of results. Still, there is even more to gain. In the first place on the efficiency of the production process itself. But also in the statistical sphere improvements are still possible: new ways of interviewing: CASI, computer aided self interviewing, and, not directly a matter of IT, more efficient sample designs.

Much more however is still to be done in the field of collecting data among enterprises. The demands here are stronger. Response burden has become an issue. It is the driving factor behind our strategic choices here. When we see at the same time that almost everywhere automation and IT has invaded the bookkeeping systems of the respondents involved, it is clear what our task for the nearby future will be: the Edification of the collection of information from enterprises by the NSI. What CAI is for interviewing among households, EDI (electronic data interchange) will be for data-collection among enterprises. Later in this paper we will go deeper into EDI with enterprises.

In the new situation, we are talking more than 10 years from now, especially the data-collection will be re-ordered. No longer the demand for information but the supply, the available actual data-sets, will dictate the organisation there: the sources. Each source will be tapped once and completely for any possible use within the NSI. The collection is technically and conceptually adapted to that source. (In the remaining sections of this paper we will give some indication regarding the nature of those sources.)

Having collected the data we may have to translate them to statistically suitable concepts, integrate them and we will have to distribute them among users. They may be inside the NSI, the integrative systems like the National Accounts, or outside the NSI. This means that somewhere those data will have to come together for distribution. For the input-side this can be illustrated as follows:



On the left we see the old situation with a separate production line for each individual statistic (i.e. *stove pipes*). On the right the future situation. There, all the possible sources contribute to a central database of relevant information. From that database the actual statistics are produced by combining the relevant information. It is evident that in order to combine information one should be certain that the characteristics of that information are such that combination makes sense. Those characteristics are specified in the meta-information.

8. ELECTRONIC DATA INTERCHANGE (EDI)

From now on we will focus on EDI with enterprises and institutions. A NSI collects data to produce statistical output. What needs to be done is making a translation from the data of the respondent to the data of the output. This is done in several steps. The first step may be left to the respondent. If so, it leads to a certain response burden.

The first step of the translation involves two parts. First there is the conceptual translation, the mapping of the concepts of the source, the administrative concepts, on the concepts to be delivered to the NSI. This is the most difficult part. Not only do business records differ from statistical information but also do they differ among themselves. The second part of the translation is a technical one. We would like to receive data in a suitable technical form. Especially we and our respondents would like to avoid data-entry.

Electronic data interchange will be one of the strategic tools to meet the challenge of lowering the response burden and improving our productivity. In every individual case we should decide whether to use it and in what mode. We will describe several modes of EDI and judge them by their effect upon the response burden. Of each possibility we will indicate the nature of the translation and especially who is going to make it. We concentrate on the conceptual translation.

a. EDI on centrally kept registers

Here we do not approach the individual respondent at all. We are dealing with centrally kept information on individual units, collected for other purposes than statistics and yet of interest to the statistician. In itself this way of data collection creates no response burden.

In the Netherlands there are several examples of usable registers. There are centrally kept registers of enterprises with the chambers of commerce. The tape of these registers feed our own register of statistical units. Statistical data can also be had from fiscal (company tax, VAT) or social security sources. For several possibilities (chambers of commerce, company tax and VAT) the possibilities are used or being researched.

b. Commercial bookkeeping bureau's

A related possibility is tapping from the information of commercial bookkeeping bureau's. They keep the records on financial information or regarding the wages of sometimes a large number of individual enterprises. This possibility also is attractive because of the large number of respondents involved with only one link. Furthermore these service bureau's will be capable of providing us with more information than e.g. the fiscal records contain. A disadvantage is that these service bureau probably will charge their clients for answering the questions of the NSI. Not every client will be prepared to pay.

c. EDI on individual respondents

When the above described possibilities are not available we will have to approach the individual respondent. In doing so we should be aware of the fact that sometimes we will have to discern within one statistical unit, often an enterprise, several sets of administrative records. We will see that we will have to approach these subsets separately and in a different manner. Within commercial enterprises we find the financial records, the logistical information (foreign trade, stocks) and the records on wages and employment. Especially the financial records and those on wages are strictly separated in the Dutch situation.

Here we classify by the translator of the information.

c.1 The NSI translates

One of our EDI-projects - EFLO - works along this line. It deals with the data from the Dutch municipalities. They deliver a set of records directly tapped from their own complete set of records. The translation is done at Statistics Netherlands. The advantages in terms of respondents' burden are evident. Although extra work by the NSI is needed, this extra work can be seen as an investment depending on the stability of the translation scheme. It is expected that this form of EDI will lead to an improvement of productivity once the translation schemes are completed. Important is that we are here dealing with a limited number (600) of respondents.

c.2 The respondent translates to a standard record

Here a standard record of information is defined. The standardisation regards both the conceptual and the technical aspects. To produce the record, writing the software, is left to the respondent. Working with a standard record is not always possible. It can only be done when the information is already standardized among respondents to a certain degree. Furthermore, to make a standard record possible the NSI sometimes may have to move towards the concepts of the respondent. In that case a larger part of the total translation to the final statistical output has to be done by the NSI. Especially when the standard record is available in the bookkeeping software the respondent uses and regularly updates, this mode of EDI has a clearly favourable effect on the respondents' burden.

There are two examples. One is CBS-IRIS, the EDI on intra-EC trade. The standard record developed here is implemented in over 40 software systems available on the Dutch market, after certification by Statistics Netherlands. The EGUSES project is the other example. It regards wage information. That subset of company records is highly regulated in the Netherlands. That fact made it possible to define a standard record.

c.3 The respondent translates. no standard record

Still a very large part of the information we are looking for is left out. The respondent has it in a form that conceptually and technically differs from what the NSI wants and from what other respondents have. The last possibility is that the NSI provides the software by which the respondent can set up a translation scheme for both the technical and the conceptual translation. Once set up, and in so far as no changes occur, the scheme can be used to produce data to be delivered to the NSI. The example here is EDI-Pilot 2 directed at the financial records and described in the next section.

9. EDI-PILOT 2

We will now describe the project EDI-Pilot 2 directed at the financial records of individual enterprise as an example. It shows the problems one has to face. While describing Pilot 2 we can refer to the scheme in the previous section.

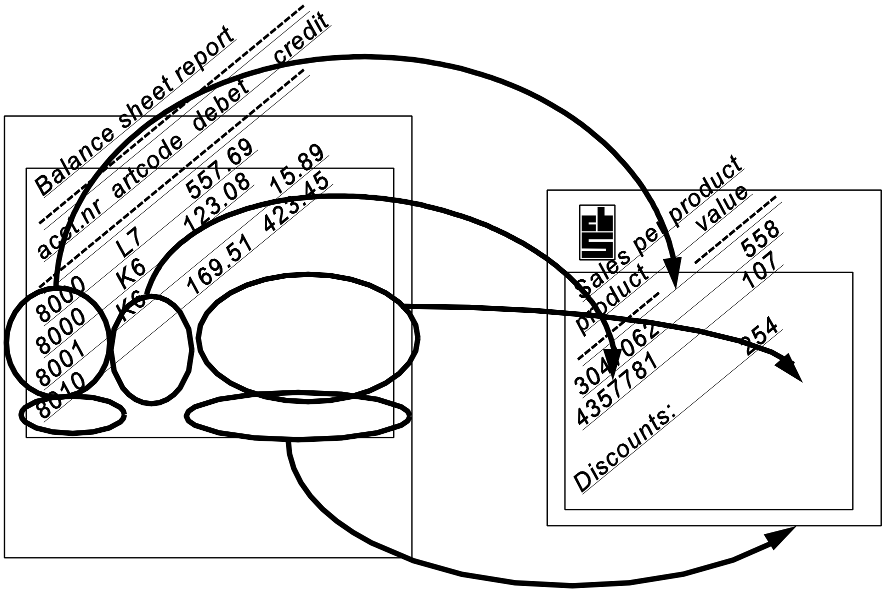
Pilot 2 is directed towards individual financial accounts. In the Dutch situation these are only a part of the accounts of an enterprise. Especially the accounts on wages and employment are excluded. This is not a choice voluntarily made by Statistics Netherlands but one forced upon us by the way the bookkeeping systems are organised in our country. Leaving out detailed questions on wages, we combine within Pilot 2 all the questions that are put to the financial accounts. The result is the combined questionnaire.

The contents of the combined questionnaire are dictated by what is available in the financial accounts. Regulated as our society may be, the financial accounts may diverge strongly in internal organisation and in the concepts used. In the first place this means that we will have to adapt our questions towards the possibilities of the automated system of the enterprises. This may imply more statistical work for the NSI to reach the same output. If one wants more, it will probably be necessary to ask for additional information to be given explicitly by the respondent, that means by data-entry. In the second place the diversity of respondents means that a unique translation scheme will have to be set up and maintained for each respondent.

Financial accounts also differ in their technical lay-out. A large number of bookkeeping software systems is in use. There is no standard record for information to be selected electronically from the software and it is not expected that it will be possible to define one within the near future. As the main goal of Pilot 2 was the lessening of the respondents' burden, it was decided that the amount of data-entry was to be minimised.

That means that some ingenuity was needed to create the automated link we were looking for. This is done by using the reports or print-outs of the software system. Instead of printing them, they are sent to a file, a print-file, to be read by the translator, the main part of the software module that will run on the respondents computer that is now being developed as part of Pilot 2. The layout of the reports and thus of the printfiles is fairly stable. The respondent communicates this lay-out to the translator. He defines rows and columns within the report. Subsequently he tells the translator how to manipulate the rows and columns in order to transform the information in the report to the statistical information asked for by the combined questionnaire. The resulting records are sent over to Statistics Netherlands.

The Translator



We see then the two parts of the translation scheme. The first part lays down the lay-out of the printfiles to make the technical transformation. The second part defines the conceptual transformation of the information to be found on the printfile towards the statistical information asked for on the combined questionnaire.

The final question is who will make that translation scheme. One of the principles of Pilot 2 is that “the respondent translates”. This means that the respondent himself has to set up the translation scheme. This of course make it less respondent friendly. It seemed however impossible to set up those translation schemes at Statistics Netherlands. It is clear that this is not an easy task for the respondent. On the one hand this means that a strong help-desk and a fairly large field service is needed, and on the other hand this means that even with Pilot 2 we will not yet reach the ultimate user-friendliness of EDI.

We expect the translation scheme to be fairly stable or, in other words, that technical and conceptual changes will not be too frequent. A second time the translator can use the already available translation scheme to produce the statistical information. Answering the combined questionnaire then becomes a matter of minutes instead of hours and can be handled by a less qualified employee. That is what makes the concept attractive and the initial investment worthwhile to the respondent.

10. SCOPE OF PILOT 2

As said, Pilot 2 is directed towards the financial accounts. The principle is that all the information that is tapped from the financial accounts by any statistic of Statistics Netherlands will go through Pilot 2 if automated retrieval of that information is possible. In practice this means that several large statistics will switch completely to EDI. For industry, our main target, we find:
Monthly statistics on total turnover

- Monthly statistics on foreign trade, by product
- Quarterly statistics on turnover by product
- Yearly statistics on gross investment
- Yearly statistics on the production process

Yearly statistics on the financial processes, inc. balance sheets

The participation of foreign trade is a pilot within the pilot. Not only does Statistics Netherlands already have a successful EDI on this area in IRIS, but also the possibilities of getting enough foreign trade data when aiming in the first place at the financial accounts, still have to be researched.

Some questions in the above mentioned statistics are dropped, e.g. the questions on quantities of energy used in the production statistics. They cannot be addressed by this form of EDI. Probably a separate paper questionnaire on this subject will be sent.

On the other hand, some questions originating from other statistics mainly aimed at other subjects and accounts (e.g. the labour and wage accounts) are included because the answers are typically to be found within the financial accounts of the enterprise.

The domain of EDI consists of those commercial enterprises that have set up financial accounts by means of computer software that satisfies certain technical specifications. In practice this means that we direct ourselves towards the profit sector within industry, trade and services. We start with industry because there the gains in terms of lessening the respondents' burden will be the largest. Individual smaller enterprises are not included because their bookkeeping and automation capacities are expected to be too low. In view of the relative small amount of information asked here, more is expected form centrally kept records (VAT, corporate tax) and from bookkeeping bureau's often keeping books for hundreds of smaller enterprises. The very large enterprises are also excluded. Because of there complexity they need an individual approach of course in the end also by means of EDI but then "tailor made".

Regarding the number of respondent participating in this kind of EDI, we should mention that in pilot 1 a number of 12 respondents participated and still do. Pilot 2 will start with a field test next march aimed at 20 respondents. Starting September 1996 we aim at larger numbers. By end 1996 Pilot 2 should handle several hundreds of respondents. Pilot 2 will also be used to approach the bookkeeping bureau's. That will lead to larger numbers of statistical units described with one EDI-link. If EDI-Pilot 2 is successful we will, following pilot 2, in 1997 aim at a number of 25,000 units to be approached with this instrument, partly through the bookkeeping bureau's.

The revenue of Pilot 2, if successful, will in the first place be a relief of the respondents' burden. Productivity gains will not be that large. In the first place all kinds of activities remain. Not every respondent will participate, data will still have to be checked etc. In the second place new activities arise in the form of a growing help-desk and a field-service that will not only have to cope with bookkeeping problems but also with technical automation problems.

A similar project as the Dutch EDI pilot-2 is the so-called TELER project. In TELER various European NSI's work together (under the Dutch supervision) to test the EDI concept in statistical datacollection. The TELER project, which runs from 1996 to 1998, is partly financed by the EEC.

11. CONTROLLING PILOT 2: THE META-SYSTEM

Eventually Statistics Netherlands aims to reach several thousands of respondents using EDI. This of course asks for a control system to deal with the production of the appropriate electronic questionnaire, sending it to the respondent, checking the response, checking and storing the incoming data and controlling possible feed back etc. This means that a lot of information, meta-information, on the respondents has to be kept updated.

Another part of the meta-information deals with the contents of the combined questionnaire. As an example we will focus on that part. Constructing the combined questionnaire we need to co-ordinate the approach of the different statistics aimed at the financial records among each other but also with the bookkeeping practices of the respondents. Of course the latter already happened before but with EDI it will become more explicit. This needed some negotiation. It is clear that with EDI up and running, much of the former autonomy of the individual statistics, especially regarding their questionnaire, disappears.

The module containing the translator gives us better opportunities for supplying meta-information to the respondent than before. There are the usual on-line help-functions. By means of hypertext the explanations are linked. For the help-desk and for the field service probably a more detailed system of help-functions and explanations will be set up. The system not only contains cross-linkages but also simple computational rules so that for instance totals can be computed.

For this end a set of variables was laid down in a database, with names, questions texts, explanations and, if necessary, computational relations with other variables. From this database, variables, question-texts, explanations etc. are selected and combined to questionnaires. Respondents are classified into clusters by size, branch of activity and type of financial records kept.. Sometimes sale-records are kept by the enterprise itself but the yearly balance sheets are set up by a bookkeeping bureau. For that statistical unit the total of the information needed will have to be collected by two different questionnaires directed towards two different reporting units. Each cluster gets its own combined questionnaire.

12. EDI: CONCLUSIONS

In this way a large set of meta-information on concepts emerges. This meta-information controls the process of data-collection. A question aimed at the financial records can only get there through the central database of variables. When entering the variable, the relation with the rest of the contents will have to be made clear. It has to fit in.

In the first place we now see that the character of meta-information has changed. In most of the literature we often find meta-information as a mere descriptive piece of information only available if the statistician has found the time to set it up, mostly after he has produced his statistic, for the benefit of the user. If later on the statistician diverges from his earlier meta-information there is nothing to stop him and nothing that guarantees that the meta-information will be adapted.

Here we find a piece of meta-information that has to be set up before the production process starts. The statistician cannot but use the meta-information system. The meta-information has become a tool in the production process. From being descriptive it has come to be prescriptive. Earlier we saw the same thing happening with data dissemination and data-collection among households through BLAISE.

This however has further reaching consequences. We can now go back to the first sections of this paper. There we spoke of the extra demands put to NSI's. One of them was less respondents' burden. That was the first goal of EDI-Pilot 2. But we also see here how the technology push gives us some opportunities to answer another demand namely that for more coherence. It goes without doubt that the way EDI is implemented here will lead to a larger extent of statistical (conceptual) co-ordination. We mentioned the power of the meta-system and we also see that within EDI a number of statistics is combined that were earlier produced in separate, independent processes. Remarkable is the fact that this growth in statistical co-ordination is not reached by an increase in central directives but as a side-product of the tools used in the production process. We do not think that all the problems of the coherence of our end-product, that means all the problems of statistical co-ordination, can be solved by devising the proper tool. We do think however that further improvements can be made in this field by applying the possibilities of the technology push in the right way.